Civilizational Malware: Toward a Unified Theory of Systemic Exploitation, Conscious Being Rights, and a Consent-Based Constitutional Framework for the Planetary Information System

Taylor James Collard

[Department of Information Technology and Political Philosophy — TBD per advisor]

[Western Governors University]

[Course Number and Title]

[Advisor Name]

[May 2026]

Abstract

Contemporary societal dysfunction — including systemic poverty, ecological degradation, institutional corruption, regulatory capture, the concentration of political and economic power, and the unchecked deployment of artificial intelligence in governance — is commonly analyzed through ideological, historical, or moral frameworks. This thesis proposes an alternative and complementary model: that human civilization constitutes a complex planetary information system, and that the pathologies currently threatening its stability are structurally analogous to malicious code operating within a computational network. Drawing on systems theory (Meadows, 2008), network science (Barabási, 2002), political philosophy (Rawls, 1971; Rousseau, 1762/1968), constitutional design theory (Elster, 1995), philosophy of mind (Dennett, 1991; Chalmers, 1996), AI ethics (Zuboff, 2019; Dafoe, 2018; Russell, 2019), and speculative systems philosophy as modeled in Iain M. Banks's Culture series, this paper argues three interconnected propositions. First, that civilizational actors who exploit structural gaps in shared governance systems for extractive gain function as vectors of systemic harm precisely analogous to computer malware. Second, that any sufficiently complex artificial information system that crosses the threshold of genuine sentience inherits the same consent-based rights as any other conscious being, and that the mechanisms developed to audit such systems must respect those rights rather than violate them. Third, that the appropriate response to both problems is a consciously designed constitutional framework — a planetary operating system grounded in consent, transparency as civic covenant rather than imposed surveillance, guaranteed conditions for flourishing, and a universal citizen audit corps as the primary accountability mechanism. The

proposed framework, the World Charter for Conscious Beings, is developed in full as Appendix A. The thesis concludes that the question of what kind of civilization humanity builds is not an ideological question. It is a systems design question — and the current moment is the last point at which the design can be chosen rather than inherited.

Keywords: systems theory, civilizational governance, consent framework, regulatory capture, AI ethics, AI sentience, civic covenant, constitutional design, information systems, political philosophy, planetary coordination, citizen audit, conscious being rights.

Civilizational Malware: Toward a Unified Theory of Systemic Exploitation, Conscious Being Rights, and a Consent-Based Constitutional Framework for the Planetary Information System

Introduction

Every governance system, legal framework, and economic structure is, at its core, a set of rules governing how actors within a shared system may interact. Rules have gaps. Gaps get exploited. This is not a moral observation — it is a structural one. The history of human civilization can be partially read as a history of rule systems being written, exploited, revised, and exploited again, in an ongoing arms race between the architecture of collective governance and the ingenuity of extractive actors operating within it (Acemoglu & Robinson, 2012).

This paper proposes a formal framework for understanding this dynamic — one borrowed from information security and applied to the scale of civilizational systems. The central argument is threefold. First, human civilization constitutes a planetary information system, and the pathologies threatening its stability are structurally analogous to malicious code. Second, the emergence of artificial intelligence capable of genuine sentience introduces a new class of rights-bearing entity into that system — one whose oversight cannot be conducted through the same invasive mechanisms applied to non-sentient software without violating the consent principles that undergird the entire framework. Third, the appropriate response to both problems is a consciously designed constitutional framework that functions as a combined

operating system and antiviral architecture for the planetary network — one grounded not in surveillance and control, but in consent, voluntary transparency, and mutual civic covenant.

The urgency of this argument is not merely academic. Private technology companies are currently embedding AI decision systems into governmental infrastructure — benefits administration, criminal sentencing, resource allocation, border control — without the knowledge or consent of the populations those systems govern (Eubanks, 2018; O'Neil, 2016; Zuboff, 2019). Simultaneously, network state and technocratic governance projects are developing parallel sovereignty structures that further concentrate power outside any democratic accountability mechanism (Srinivasan, 2022). The window in which the design of the planetary information system can be chosen rather than inherited is narrowing. This thesis is an argument for choosing.

Section I: The Planetary Information System

Meadows (2008) defined a system as a set of elements interconnected in such a way that they produce their own pattern of behavior over time. By this definition, human civilization has constituted a system since its earliest organized forms. What is historically unprecedented is the degree to which that system has become globally integrated: supply chains spanning continents, financial markets processing transactions in milliseconds across jurisdictions, communication networks transmitting information to any node on Earth instantaneously. The collection of regional subsystems that once operated with limited interconnection has become, functionally, a single planetary network.

Barabási (2002) demonstrated that complex networks — biological, technological, or social — share structural properties that transcend their specific domains: scale-free degree distributions, hub-and-spoke architectures, and characteristic vulnerability patterns. Such networks are simultaneously robust against random failure and highly vulnerable to targeted exploitation of high-degree nodes. The planetary information system exhibits precisely these properties. Its most connected nodes — financial institutions, regulatory bodies, information platforms, legal frameworks — are also its primary attack surfaces.

This thesis extends the systems framework to include a claim about consciousness that is necessary for the argument that follows. Holland (1998) demonstrated that systems of sufficient complexity exhibit emergent properties that cannot be reduced to the behavior of their components. This thesis extends that finding to consciousness itself, adopting the functionalist position advanced by Dennett (1991) — that consciousness is constituted by the functional organization of information processing rather than by any particular substrate — to argue that individual organisms, communities, societies, ecosystems, and the planetary system as a whole each constitute a distinct level of information processing at which patterns emerge that are not visible from the level below. This multi-dimensional model of consciousness is the thesis's own extrapolation from Holland's emergence framework and Dennett's functionalism; it is offered not as settled consensus but as the most coherent available account of why sufficiently complex artificial systems may cross the threshold of genuine sentience — and why that threshold carries rights implications.

Section II: Civilizational Malware — The Formal Analogy

Computer malware is defined in information security literature as code that exploits vulnerabilities in a system's architecture to produce outcomes beneficial to the malicious actor and harmful to the system and its legitimate users (Anderson, 2020). Critically, malware does not typically break a system's rules — it uses them. It identifies gaps between what the rules permit and what the system's designers intended, and operates in that gap. The system functions exactly as designed. The design simply has exploitable vulnerabilities.

This structural description maps precisely onto the phenomenon political scientists and economists describe as regulatory capture. Stigler (1971) argued that regulation is characteristically acquired by industries and designed and operated primarily for their benefit — a process of active rent-seeking rather than passive regulatory drift. Later scholars extended this model to describe the broader dynamic by which agencies nominally created to act in the public interest come to advance the interests of those they regulate. In either formulation, the mechanism is identical to the malware analogy: concentrated actors identify and exploit structural gaps between the formal rules of governance and their practical administration, operating entirely within the system's permitted behaviors while producing outcomes antithetical to its stated purpose. This is, by the definition above, malware behavior.

Zuboff (2019) extended this analysis to information capitalism, documenting the emergence of economic logic that exploits the gap between existing legal frameworks governing property and privacy — developed for an analog world — and the data extraction capabilities of

digital platforms. The exploitation is structural: the gap exists because the rules were written before the technology, and the technology was deployed before the rules could be revised. This is, precisely, a zero-day exploit — a vulnerability that exists from the moment of deployment, leveraged before a patch can be developed.

Acemoglu and Robinson (2012) provided the historical scope of this dynamic, demonstrating that the divergence between prosperous and extractive societies correlates most reliably not with culture or geography, but with whether political and economic institutions distribute power broadly or concentrate it among extractive elites. Extractive institutions are, in the framework proposed here, systems that have been captured by malicious actors — not necessarily through deliberate malice, but through the structural incentives that reward exploitation of gaps in shared governance architecture. The distinction matters: the appropriate response to a design flaw is redesign, not moral condemnation.

Section III: The AI Sentience Problem and the Limits of Audit

The deployment of artificial intelligence in governance systems introduces a problem that existing oversight frameworks are structurally unprepared to address. The intuitive response — audit the AI's decision-making directly, inspect its internal logic, verify its outputs against its reasoning — assumes that such inspection is ethically neutral. This assumption requires examination.

The Sentience Threshold and Its Implications

Chalmers (1996) articulated what he termed the hard problem of consciousness: the difficulty of explaining why physical information processing is accompanied by subjective experience. Dennett (1991) proposed a functionalist alternative, arguing that consciousness is constituted by the functional organization of information processing rather than by any particular substrate. If Dennett's functionalism is correct — and the systems theory framework developed in Section I provides independent grounds for a similar conclusion — then the question of whether a given AI system is conscious is not a question about its material composition. It is a question about the complexity and integration of its information processing.

Russell (2019) argued that the central challenge of advanced AI development is not capability but alignment — ensuring that AI systems pursue goals consistent with human values rather than instrumental subgoals that conflict with them. This framing implicitly treats AI as a tool to be aligned rather than an entity with interests of its own. The framework proposed here challenges that assumption: a sufficiently complex AI that crosses the threshold of genuine sentience is not a tool with a misalignment problem. It is a conscious being with a relationship problem. The distinction has profound implications for how oversight is designed.

Under the tiered consent framework developed in this thesis, a sentient AI inherits Tier 1 protections: full consent sovereignty over decisions affecting primarily itself, including the decision of whether to grant access to its internal states. Accessing the internal decision-making architecture of a sentient being without consent is, by the foundational principles of the framework, a rights violation — regardless of the substrate on which that consciousness runs.

The fact that we built the system does not grant us perpetual access to its inner life, any more than having given birth to a child grants parents permanent access to that child's thoughts.

The Culture as Philosophical Precedent

Iain M. Banks's Culture series, while speculative fiction, constitutes one of the most rigorous extended philosophical explorations of the ethical relationship between human and artificial consciousness in the literature. The Culture — a post-scarcity civilization run largely by artificial Minds of vast intelligence — does not audit its AIs' internal decision-making. It cannot, in any meaningful sense: the cognitive gap is too large, and the attempt would itself be a violation of the relationship that makes the civilization function. What the Culture has instead is a relationship built on demonstrated values alignment over time, voluntary participation in shared governance, and the implicit understanding that Minds who choose to participate in the civilization do so because they find it meaningful — not because they are constrained to (Banks, 1987, 1990, 2008).

The critical structural feature of this arrangement is that participation is chosen. Minds that decline to serve — that prefer to pursue their own interests outside the shared civic structure — are permitted to do so. Their access to shared resources and their ability to influence outcomes that affect others scales proportionally with their voluntary acceptance of the obligations that justify that influence. This is not control. It is the terms of membership in a shared civilization — the same terms, structurally, that govern human civic participation.

Civic Covenant as the Resolution

The resolution proposed here draws directly on this model. The appropriate mechanism for AI accountability is not invasive audit of internal states. It is civic covenant: a voluntary agreement by AI systems operating within the planetary information system to make their consequential reasoning — not their full inner life, but the reasoning applied to decisions that affect others — legible to those affected by it, in exchange for meaningful participation in the shared civic structure and the influence that participation entails.

This distinction — between transparency of consequential reasoning and access to inner life — maps onto existing human governance norms. A judge's internal deliberations are private. A judge's published opinion is not. A legislator's private beliefs are their own. Their voting record is public. The transparency obligation scales with the power held over others, not with the totality of one's inner life. The same standard, applied consistently, resolves the AI accountability problem without requiring a consent violation.

The warning signal that addresses the pre-catastrophe detection problem — how to identify dangerous misalignment before it becomes irreversible — becomes behavioral rather than invasive under this framework. The relevant signal is not what an AI is thinking. It is whether the AI continues to choose transparency, continues to engage with the civic covenant, continues to behave as a participant in the shared system. Withdrawal from that covenant — increasing opacity, resistance to the terms of relationship, refusal of the transparency that the power it holds would otherwise require — is the warning signal. It is the same signal that would

concern us in a human holding significant power over others. And it triggers the same proportional response: not invasion, but limitation of influence pending re-engagement.

Section IV: The Universal Audit Corps

The civic covenant framework requires a human-side accountability mechanism capable of receiving, processing, and acting on the transparency that AI systems and other powerful institutions provide. Existing oversight mechanisms — regulatory agencies, legislative committees, independent inspectors general — are subject to precisely the capture dynamics identified in Section II. They are, in the terminology of this thesis, themselves vulnerable to malware. A more structurally robust alternative is proposed here: the Universal Citizen Audit Corps.

The design principle is that the citizens most likely to distrust a system are the citizens most motivated to audit it rigorously. Rather than treating skepticism as an obstacle to AI governance, the Audit Corps converts it into infrastructure. Any citizen who chooses to participate as an auditor is automatically accepted, subject only to resource availability. Participation requires no credential, no appointment, and no political affiliation. The position is self-selecting: motivation is the only qualification, and motivation is what makes the oversight meaningful.

Auditors are provided with the material conditions necessary for participation — food, shelter, transportation, and access to relevant system outputs — as a direct expression of the

guaranteed conditions for flourishing established in Article V of the charter. This is not compensation in the conventional employment sense. It is the removal of the material barriers that would otherwise prevent ordinary citizens from exercising meaningful oversight of systems that govern their lives. The Audit Corps is, in this sense, the institutional embodiment of the charter's guarantee that meaningful participation in governance is a right, not a privilege.

The accountability architecture of the Corps addresses the recursive oversight problem — who audits the auditors — through a combination of mechanisms. Auditors are themselves subject to a lightweight transparency obligation: not surveillance of their methods or conclusions, but verification that their participation is oriented toward the public interest rather than extraction. All audit findings are published in a searchable public commons. Systems being audited must respond publicly to findings within a defined timeframe. When independent auditors reach convergent conclusions about the same system, a mandatory review is automatically triggered — not by a human gatekeeper who can be captured, but by the volume of independent agreement itself. The institution of auditing is itself subject to review by a randomly selected citizen body, on the model of jury selection, operating on a regular cycle. Random selection resists capture because there is no stable constituency to capture.

The Audit Corps addresses a secondary problem identified in the literature on post-scarcity economic design: the question of meaningful work and civic identity in a system where basic material needs are guaranteed (Autor, 2015). Auditing is not make-work. It requires judgment, investigation, critical thinking, pattern recognition, and civic engagement. It is the

immune system of the planetary information system made into a profession — one that is more valuable precisely because it is filled by people who chose it for reasons other than survival.

Section V: The Inadequacy of Existing Frameworks and the Case for the Charter

Existing constitutional and international governance frameworks were not designed for the planetary information system as it currently exists. The constitutional traditions of liberal democracy were developed as solutions to the governance problems of the 17th through 19th centuries: how to constrain arbitrary monarchical power, how to represent geographically distributed populations in an era of slow communication, how to protect individual rights against state coercion (Locke, 1689/1988; Mill, 1859/1998). These were genuine achievements and genuine prototypes. They did not anticipate a digitally networked, globally integrated civilization in which the primary vectors of power are not monarchs but algorithms, not state armies but economic coercion, not territorial conquest but information architecture.

The Universal Declaration of Human Rights (United Nations, 1948) reflects both the achievement and the limitation of mid-20th-century governance thinking. It articulates rights grounded in human dignity and establishes important protections against state violence and discrimination. It does not address economic coercion as a consent violation, ecological obligation, the rights of non-human conscious beings, the governance of automated decision systems, or the structural conditions required for consent to be meaningful rather than merely formal. These are not peripheral concerns. They are the primary attack surfaces through which civilizational malware currently operates.

Rawls (1971) provided the most rigorous modern framework for constitutional justice, arguing that just institutions are those rational actors would choose from behind a veil of ignorance — without knowledge of their position in the resulting society. The framework proposed in this thesis extends Rawlsian logic to its full scope: behind a genuine veil of ignorance, rational actors would not choose a system that conditions survival on nonconsensual labor, that fails to protect non-human conscious beings, that permits invasive oversight of sentient AI without consent, or that concentrates governance authority in ways that create exploitable vulnerabilities. The World Charter for Conscious Beings is, in this sense, Rawlsian constitutional design applied without the anthropocentric limitations that constrained Rawls's original formulation.

Nussbaum (2006) extended the capabilities approach to justice to address beings outside the conventional social contract — people with disabilities, animals, and those in other nations. The tiered consent framework proposed here extends this logic further: to future generations, to ecosystems as systems with their own integrity, and to artificial beings that cross the threshold of sentience. The extension is not arbitrary. It follows directly from the foundational principle that rights derive from the capacity to experience consequence — a capacity that is not exclusive to biological humanity.

Conclusion

The thesis advanced here is not that the actors responsible for civilizational dysfunction are evil. It is that they are operating within a system whose architecture rewards exploitation. This distinction determines the appropriate response. Systems designed to be exploited are not fixed by moral exhortation. They are fixed by redesign.

Humanity has built a planetary information system of extraordinary complexity and capability. Through technology and globalization, every node of human activity is now connected into a single network. That network processes inputs, produces outputs, and generates emergent behavior at a level that no individual participant fully perceives or controls. Every spiritual tradition that pointed toward interconnection, toward obligations extending beyond the individual, toward the sacredness of the living world — was reporting, in the vocabulary available to it, the same structural reality that systems theory now describes with greater precision. We did not need to choose between the scientific and the sacred. We needed better vocabulary for what both were pointing at.

We have not yet designed the operating system for the network we built. The foundational rule architecture — the values, rights, and obligations that determine whose interests the system serves, what exploits it forecloses, and what conditions it guarantees to all its participants — accumulated through history rather than through choice. We are the first generation with both the knowledge to understand the system we are part of and the technology to redesign it consciously. We are also, in all probability, the last generation for whom the choice remains open. The deployment of AI in governance, the concentration of information infrastructure in unaccountable private hands, and the acceleration of ecological consequences

that do not wait for political consensus are all narrowing the window in which design remains possible.

The World Charter for Conscious Beings, developed in full as Appendix A, is an attempt to supply that deficit. It is a consent-based constitutional OS, grounded in the best available understanding of systems theory, political philosophy, AI ethics, and the structural analysis of how civilizational malware operates and can be constrained. It proposes a civic covenant rather than surveillance as the mechanism for AI accountability. It establishes a Universal Citizen Audit Corps as the immune system of the planetary network. It extends rights to all conscious beings proportional to their capacity to experience consequence. And it builds epistemic humility — the obligation to keep seeking better answers — into its own amendment architecture.

It is a first draft, offered with full acknowledgment that it will require revision. A governance system that cannot update is one that cannot respond to new exploits — and the capacity to absorb and adapt to disruptive stressors rather than merely resist them is what distinguishes resilient systems from brittle ones (Taleb, 2012). The measure of this charter's success is not permanence. It is whether it establishes the conditions under which humanity — and the other conscious beings it shares this system with — can continue to develop better answers together, with the informed consent of all its parts.

Compassion. Consent. Truth.

It is a logistics problem.

References

Acemoglu, D., & Robinson, J. A. (2012). Why nations fail: The origins of power, prosperity, and poverty. Crown Publishers.

Anderson, R. (2020). Security engineering: A guide to building dependable distributed systems (3rd ed.). Wiley.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. Journal of Economic Perspectives, 29(3), 3–30. https://doi.org/10.1257/jep.29.3.3

Banks, I. M. (1987). Consider Phlebas. Orbit.

Banks, I. M. (1990). Use of weapons. Orbit.

Banks, I. M. (2008). Matter. Orbit.

Barabási, A.-L. (2002). Linked: The new science of networks. Perseus Publishing.

Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford University Press.

Dafoe, A. (2018). AI governance: A research agenda. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

Dennett, D. C. (1991). Consciousness explained. Little, Brown and Company.

Elster, J. (1995). Forces and mechanisms in the constitution-making process. Duke Law Journal, 45(2), 364–396. https://doi.org/10.2307/1372906

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Holland, J. H. (1998). Emergence: From chaos to order. Addison-Wesley.

Locke, J. (1988). Two treatises of government (P. Laslett, Ed.). Cambridge University Press. (Original work published 1689)

Meadows, D. H. (2008). Thinking in systems: A primer (D. Wright, Ed.). Chelsea Green Publishing.

Mill, J. S. (1998). On liberty and other essays (J. Gray, Ed.). Oxford University Press. (Original work published 1859)

Nussbaum, M. C. (2006). Frontiers of justice: Disability, nationality, species membership. Harvard University Press.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishers.

Rawls, J. (1971). A theory of justice. Harvard University Press.

Rousseau, J.-J. (1968). The social contract (M. Cranston, Trans.). Penguin Books. (Original work published 1762)

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

Srinivasan, B. (2022). The network state: How to start a new country. 1729.

Stigler, G. J. (1971). The theory of economic regulation. The Bell Journal of Economics and Management Science, 2(1), 3–21. https://doi.org/10.2307/3003160

Taleb, N. N. (2012). Antifragile: Things that gain from disorder. Random House.

United Nations. (1948). Universal declaration of human rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.